O'REILLY®

Data Algorithms with Spark

Apache Spark's speed, ease of use, sophisticated analytics, and multilanguage support makes practical knowledge of this cluster-computing framework a required skill for data engineers and data scientists. With this hands-on guide, anyone looking for an introduction to Spark will learn practical algorithms and examples using PySpark.

In each chapter, author Mahmoud Parsian shows you how to solve a data problem with a set of Spark transformations and algorithms. You'll learn how to tackle problems involving ETL, design patterns, machine learning algorithms, data partitioning, and genomics analysis. Each detailed recipe includes PySpark algorithms using the PySpark driver and shell script.

With this book, you will:

- **Learn how to select Spark transformations for optimized solutions**
- **Explore powerful transformations and reductions including reduceByKey(), combineByKey(), and mapPartitions()**
- **Understand data partitioning for optimized queries**
- **Build and apply a model using PySpark design patterns**
- **Apply motif-finding algorithms to graph data**
- **Analyze graph data by using the GraphFrames API**
- **Apply PySpark algorithms to clinical and genomics data**
- **Learn how to use and apply feature engineering in ML algorithms**
- **Understand and use practical data design patterns**

"This book is a great resource for both readers looking to implement existing algorithms in a scalable fashion and readers who are developing new, custom algorithms using Spark."

—Matei Zaharia
Asst. Professor of Computer Science, Stanford;
Chief Technologist, Databricks;
Original Creator of Apache Spark

Mahmoud Parsian, PhD in computer science, is a practicing software professional with 30 years of experience as a developer, designer, architect, and author. Over the past 15 years, he's been involved in Java server-side computing, databases, MapReduce, Spark, PySpark, and distributed computing. Dr. Parsian leads Illumina's Big Data team, which focuses on large-scale genome analytics and distributed computing using Spark and PySpark. Dr. Parsian also teaches machine learning and big data modeling and analytics at Santa Clara University.

DATA

US $69.99 CAN $87.99
ISBN: 978-1-492-08238-5

Twitter: @oreillymedia
linkedin.com/company/oreilly-media
youtube.com/oreillymedia

Data Algorithms with Spark

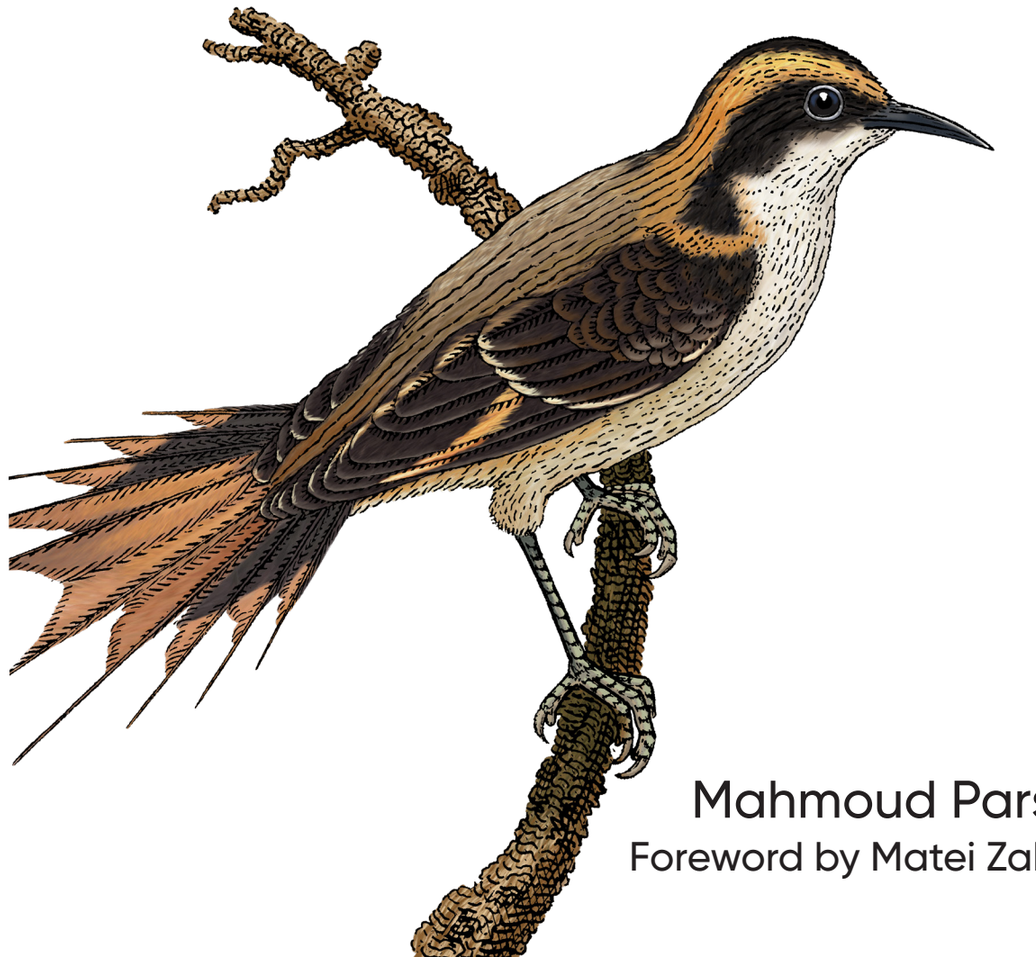
Parsian

O'REILLY®

O'REILLY®

Data Algorithms with Spark

Recipes and Design Patterns for Scaling Up Using PySpark



Mahmoud Parsian
Foreword by Matei Zaharia